

On the optimality of universal classifiers for finite-length individual test sequences.

Jacob Ziv

Department of Electrical Engineering
Technion-Israel Institute of Technology
Haifa 32000, Israel

October 8, 2009

Abstract

It has been demonstrated that if two individual sequences are independent realizations of two finite-order, finite alphabet, stationary Markov processes, an empirical divergence measure (ZMM) that is based on cross-parsing of one sequence relative to the second one converges to the relative entropy almost surely. This leads to a realization of an empirical, linear complexity universal classifier which is asymptotically optimal in the sense that the probability of classification error vanishes as the length of the sequence tends to infinity if the KL-divergence between the two processes is positive. It is demonstrated that a version of the ZMM is not only asymptotically optimal as the length of the sequences tends to infinity, but is also essentially-optimal for a class of finite-length sequences that are realizations of finite-alphabet, vanishing memory processes with positive transitions in the sense that the probability of classification error vanishes if the length of the sequences is larger than some positive integer N_0 and leads to an asymptotically optimal classification algorithm. At the same time no universal classifier can yield an efficient discrimination between any two distinct processes in this class, if the length of the two sequences N is such that $\log N < \log N_0$, even if the KL-divergence between the two processes is positive. It is further demonstrated that not every asymptotically optimal universal classification algorithm is also essentially optimal.

A variable length (VL) divergence that converges to the KL-divergence when the length of the sequences tends to infinity, is defined. Another universal classification algorithm which, like ZMM is also based on cross-parsing, is shown to be optimal relative to the VL divergence (rather than being just essentially optimal) for any two finite-length sequences that are realizations of vanishing-memory processes.

Index terms : universal classification, universal data-compression.

1 Introduction, notations and definitions

A device called a **classifier** (or discriminator) observes two N -sequences whose probability laws are Q and P respectively (Q and P are defined on doubly infinite sequences in a finite alphabet \mathbf{A}). Both Q and P are unknown. The classifier's task is to decide whether $P = Q$, or P and Q are sufficiently different according to some appropriate criterion Δ . If the classifier has available an infinite amount of training data (i.e. if N is large enough), this is a simple matter. However, here we study the case where N is finite.

The results in this paper are generalization of the results in [1] for finite-length test sequences rather than infinite ones.

Consider random sequences from a finite alphabet \mathbf{A} , where $|\mathbf{A}| = A < \infty$. Denote ℓ vectors from \mathbf{A} by $z^\ell = z_1, \dots, z_\ell \in \mathbf{A}^\ell$, and use upper case Z 's to denote random variables. When the superscript is clear from the context, it will be omitted. Similarly, a substring $Z_i, \dots, Z_j; -\infty \leq i < j \leq +\infty$ is denoted by Z_i^j .

Let a class of "vanishing memory" processes M be defined as follows:

$M = M_{k_0, \beta, \ell}$ is the set of probability measures on doubly infinite sequences from the set \mathbf{A} , with the following properties:

A) Positive transitions property:

$$P(X_1 = z_1 | X_{-\infty}^0 = z_{-\infty}^0, X_2^\infty = z_2^\infty) \geq \alpha > 0$$

for all sequences of $z_{-\infty}^\infty$ for every $P \in M$.

B) Strong Mixing condition (following [2], Eq. (9)):

Let $\{X_i\}, -\infty < i < \infty$, be a random sequence with probability law $P \in M$. We further assume that $\{X_i\}$ is a stationary ergodic process where every member in M_ℓ satisfies the following condition:

Condition 1 Let $\sigma(X_i^j; -\infty \leq i, j \leq +\infty)$ be the σ -field generated by the subsequence X_i^j . Then, there exists an integer k_0 , such that for all $k \geq k_0$, all $A \in \sigma(X_{-\infty}^0)$ and all $B \in \sigma(X_k^\infty)$

$$\frac{1}{\beta} \leq \frac{P(B)}{P(B|A)} \leq \beta \tag{1}$$

for $P(A), P(B) > 0$ and $\beta \geq 1$.

C) $P[X_1^N : P[X_1^\ell] < 2^{-\ell R}] \leq \alpha$ for every $P \in M_\ell = M(\alpha, \beta, k_0, \ell)$.

The constants k_0, β, R and ℓ do not depend on P .

The condition in B) is reminiscent of ϕ -mixing but is not identical to it. We remark that if P is any irreducible, aperiodic finite-order Markov process, this condition will be satisfied. Furthermore, the “positive transitions” condition may be guaranteed by dithering prior to the classification process, without violating the strong mixing condition. The condition in C) is satisfied by any ergodic process for some ℓ , by the Asymptotic Equipartition Property (AEP) of information theory.

2 Statement of results

Let the normalized N -th order K-L divergence between Q and $P \in M$ be:

$$D_N(P\|Q) = \frac{1}{N} K(P^N\|Q^N) \triangleq \frac{1}{N} \sum_{\mathbf{Z} \in \mathbf{A}^N} P(\mathbf{Z}) \log \frac{P(\mathbf{Z})}{Q(\mathbf{Z})}$$

where P^N, Q^N are the N -dimensional marginal measures of Q, P , and $K(*\|*)$ denotes the conventional Kullback-Leibler divergence. Logarithms are taken on base 2 and obey $0 \log 0 \equiv 0$.

Note that due to the positive transitions property of the collection M , $D_N(P\|Q) \leq \log \frac{1}{\delta} < \infty$ for every $Q \in M$.

The asymptotic K-L divergence between Q and P is given by:

$$D(P\|Q) = \limsup_{N \rightarrow \infty} D_N(P\|Q) \tag{2}$$

Formally, given an N -sequence \mathbf{Y} which is a realization of Q and another N -sequence \mathbf{X} which is a realization of P , we define a classifier f_c (c-for “classifier”) as a mapping of (\mathbf{X}, \mathbf{Y}) to $\{0, 1\}$,

$$f_c : \mathbf{A}^{2N} \times M \rightarrow \{0, 1\}$$

where $f_c = 1$ declares Q to be different from P , $f_c = 0$ means $Q = P$.

For any collection $\hat{M} \in M$ of probability measures $P_i; 1 \leq i \leq |\hat{M}|$, define

$$\lambda(P_i, \Delta, \hat{M}) =$$

$$P_r[(\mathbf{X}, \mathbf{Y}) : \text{either } f_c(\mathbf{X}, \mathbf{Y}) = 1 \text{ and } P_j = P_i, \text{ or}$$

$$\text{for some } P_j : D(P_i \| P_j) \geq \Delta, f_c(\mathbf{X}_i, \mathbf{Y}_j) \equiv 0] \quad (3)$$

where Δ is a fidelity criterion.

Also, let

$$\lambda(\hat{M}) = \sup_{P_i \in \hat{M}} \lambda(P_i, \Delta, \hat{M}) \quad (4)$$

We seek classifiers $f_c(\mathbf{X}, \mathbf{Y})$ which are derived from two “training sequence” \mathbf{Y} and \mathbf{X} of length N and which will make $\lambda(\hat{M})$, the classification error, small for any $\hat{M} \in M$.

A classifier f_c is said to be *asymptotically optimal* if the probability of classification error tends to zero for every $\hat{M} \in M$ as the length of the two sequences tends to infinity.

The efficiency of different universal classifiers that are asymptotically optimal should also be judged by the rate at which the the corresponding classification error tends to zero as N increases, since, after all, one has to deal with finite-length sequences.

In order to evaluate the efficiency of a universal classifier for finite-length sequences, we may consider appropriate fidelity function $F(Q(N), P(N))$ other than $F(P(N), Q(N)) = D_N(P \| Q)$, as long as it converges to the “classical” KL-fidelity function $D(P \| Q)$ as N tends to infinity.

Hence, we limit the discussion to the class \mathbf{F} of fidelity functions $F(P^N, Q^N)$ such that

$$\limsup_{N \rightarrow \infty} F(P^N, Q^N) = D(P \| Q) \quad (5)$$

almost surely, where $D(P \| Q)$ is the K-L divergence

Now, given a particular fidelity function $F \in \mathbf{F}$, and a collection $\hat{M} \in M$ of probability measures $P_i; 1 \leq i \leq |\hat{M}|$, assume that \mathbf{X} is a realization of P_i and that \mathbf{Y} is a realization of P_j , and define

$$\lambda_F(P_i, \Delta, \hat{M}) =$$

$$P_r[(\mathbf{X}, \mathbf{Y}) : \text{either } f_c(\mathbf{X}, \mathbf{Y}) = 1 \text{ and } P_j = P_i, \text{ or}$$

$$\text{for some } P_j : F(P_i, P_j) \geq \Delta, f_c(\mathbf{X}_i, \mathbf{Y}_j) \equiv 0] \quad (6)$$

where Δ is a fidelity criterion.

Also, let

$$\lambda_F(\hat{M}) = \sup_{P_i \in \hat{M}} \lambda_F(P_i, \Delta, \hat{M}) \quad (7)$$

Hence, every classifier that utilizes a fidelity function $f \in \mathbf{F}$ is asymptotically optimal.

Let us first start with the classical case where $F(P_N, Q_N) = D_N(P\|Q)$. Let $\lambda(M) = \lambda_F(M)$ for this particular fidelity function.

Following [2], it is shown in Theorem 1 below that the classification error that is associated with *any* universal classifier that has only the two sequences, \mathbf{X} and \mathbf{Y} at its disposal, is close to one, for the class M_ℓ , if $N \leq N_0 2^{-\epsilon\ell}$, where $N_0 = 2^{R\ell}$ and where R and ℓ are the parameters that define the class of processes M . But is there an optimal universal algorithm that will yield a vanishing classification error probability $\lambda(M)$ for $N \geq N_0 \epsilon\ell$?

An asymptotically optimal classifier f_c is said to be also *F – optimal* over M , for *finite* sequences if the probability of classification error $\lambda_F(M)$ becomes negligible for training sequences longer than or equal to $N_0 2^{\epsilon\ell}$, where N_0 is some positive integer such that any universal classifier will yield a probability of classification error $\lambda_F(M)$ which is close to one, if the length of the sequences $N \leq N_0 2^{-\epsilon\ell}$. The description of such an *F – optimal* universal classifier appears in Section 2 below.

Apparently, not every fidelity function $F \in \mathbf{F}$ leads to an associated *F – optimal* universal classifier.

An asymptotically optimal classifier f_c is said to be also *essentially optimal* over M , for *finite* length sequences if there exists a collection $\hat{M} \in M$ of pairs $P, Q : D_N(P\|Q) \geq \Delta$ for which the probability of classification error $\lambda(\hat{M})$ is close to one for $N \leq N_0 2^{-\epsilon\ell}$ and becomes negligible for sequences longer or equal to $N_0 2^{\epsilon\ell}$.

It should be noted in passing that if one of the probability measures Q is fully known to the classifier, if the sequence \mathbf{Y} is of length ℓ and if the fidelity criterion is $D_\ell(P\|Q) \geq \Delta$, there is indeed a classifier $f_c(\mathbf{X}, Q)$ that is essentially optimal over the whole class M and is therefore optimal. This follows from the fact that the measure P_ℓ of highly probable ℓ -vectors can be well estimated from \mathbf{X} once $N \geq N_0 2^{\epsilon\ell}$. Hence one can generate a good estimate for $D_\ell(P\|Q)$. However, if, as

in our case, Q is not known and the classification is based only on the observed vectors \mathbf{X} and \mathbf{Y} this need not be the case any more since no good empirical estimate for Q -improbable X_1^ℓ may be generated from \mathbf{Y} unless it's length becomes much larger.

It is demonstrated that a ZMM-based classifier is asymptotically optimal as well as essentially optimal relative to the fidelity function $D_N(P\|Q)$.

A common classifier is the Empirical Statistics Classifier (ESC), where

$$d(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \sum_{i=0}^{T-1} \log \frac{\hat{P}_{\mathbf{X}}(X_{i+1}^{(i+1)n})}{\hat{Q}_{\mathbf{Y}}(X_{i+1}^{(i+1)n})}$$

where $\hat{P}_{\mathbf{X}}(Z_1^n); Z_1^n \in \mathbf{A}^n$ denotes an empirically -derived estimate of the probability of n -vectors in \mathbf{X} , where $n = \delta_0 \log N$, $0 < \delta_0 \ll 1$. and where T is an integer satisfying $Tn \leq N < (T+1)n$.

Also, Let $f_c(\mathbf{X}, \mathbf{Y}) = 1$ if $d(\mathbf{X}, \mathbf{Y}) > \frac{\Delta}{2}$ and $f_c(\mathbf{X}, \mathbf{Y}) = 0$ if $d(\mathbf{X}, \mathbf{Y}) \leq \frac{\Delta}{2}$.

It follows that for the vanishing memory class of processes M , such an ESC is asymptotically optimal since,

$$\lim_{N \rightarrow \infty} [d(\mathbf{X}, \mathbf{Y}) - \frac{1}{N} [\log P(\mathbf{X}) - \log Q(\mathbf{X})]] = 0$$

in probability. However, it is demonstrated below that the ESC is NOT *essentially optimal*.

Thus, not every universal classifier which is asymptotically optimal is also essentially optimal.

In the following converse theorem it is demonstrated that no efficient classification is possible (i.e. $\lambda(M) \approx 1$) if $N \leq N_0 2^{-\epsilon \ell}$.

Let the class $\hat{M} \in M$ be the class of processes that are generated as appears in [2,p.346, Proof of Theorem 6]. Following the proof of in [2, Theorem 6], we get the following converse theorem:

Theorem 1 : Let $Q, P \in \hat{M}$ and let $N \leq 2^{\ell(R-\epsilon)}$. Then, for all $\alpha, \epsilon, \Delta > 0$ and all $R \in 0, \log A$, there exists a $\delta_0 = \delta_0(\alpha, \epsilon, \Delta, R)$ (sufficiently small) and an ℓ_0 such that for all $\ell \geq \ell_0$ any discriminator on $M(R, \alpha, \delta_0, \ell)$ with parameters N, Δ, λ for which $N \leq 2^{\ell(R-\epsilon)}$, must satisfy $\lambda(\hat{M}) > 1 - e^{-c(\beta, \delta)N}$.

Proof of Theorem 1: By Lemma A1 in [2], there exists a collection of cyclic subsets A_i of ℓ -vectors from $[0, 1]^\ell$, each of size $2^{R_0\ell}$, and where, for some $\beta_0 (0 < \beta_0 < 1/2)$ the Hamming distance between any $x \in A_i, y \in A_j; (i \neq j)$, $d_H(x, y) \geq \ell\beta_0$

Construction of \hat{M} : At time zero, choose an ℓ -vector from \mathbf{A}^ℓ with a uniform distribution on a cyclic set A_i . Repeat this ℓ -vector ν times to create a $\nu\ell$ vector.

Next, add a ν' -vector consisting of the first ν' elements in the first vector chosen. Say that ν' is uniformly distributed on $[1, \ell]$. Since the sets A_i are cyclic, any length ℓ substring of this vector belong to A_i . Thus, we have defined a random $(\nu\ell + \nu')$ vector. The process \hat{P} is the concatenation of these sequences with a random-phase uniformly distributed between 0 and $(\ell\nu - 1)$, and dithered by the additional modulo 2 of an i.i.d. "noise" vector \mathbf{W} with $P_r(W_i = 1) = \delta, P_r(W_i = 0) = 1 - \delta$ [2, page 346].

By Lemma A1 in [2] it follows that by choosing δ to be small enough, the divergence $D_\ell(P_i \| P_j); i \neq j$ (as well as $D_{N_0}(P_i \| P_j)$ and $D(P_i \| P_j)$), for any two such processes can be made arbitrarily large. At the same time, the number of processes in \hat{M}_ℓ is at least $2^{2^{(R-\epsilon)\ell}}$ while there are only 2^N X sequences to cope with \hat{M}_ℓ , and by derivation similar to those of [2, Eqs (A12) and (A12)], leading to the conclusion that $\lambda_F(M) \leq 1 - e^{-Nc(\beta_0, \delta)}$ if $N_0 \leq 2^{(R-\epsilon)\ell}$, even if the measure P_j that governs \mathbf{Y} , is given.

Section 1: A ZMM-based classifier is essentially optimal

It will now be demonstrated that a classifier which is based on a variant of ZMM [1] is *essentially optimal*. Denote by $C_{77}(X_1^N)$ the the number of phrases that are generated by the LZ77 parsing of X_1^N (see [4]). Thus, $C_{77}(X_1^N)$ denotes the number of distinct phrases that are generated by applying the parsing procedure that is associated with LZ77, where each phrase is the longest incoming string of yet unparsed letters, that appears in the previously encoded data, extended by one letter.

Also, let $C_{77}(X_1^N \| Y_1^N)$ be the number of phrases that are generated by cross-parsing of X_1^N relative to Y_1^N . Thus, $C_{77}(X_1^N \| Y_1^N)$ denotes the number of phrases that are generated by applying the parsing procedure that is associated with LZ77, where in this case each phrase is the longest incoming string of yet unparsed letters in X_1^N that appears in Y_1^N where the minimum phrase length is set to be one.

Now, following [1], given two N -sequences X_1^N and Z_1^N let,

$$d_{ZMM1}(X_1^N|Y_1^N) = \frac{1}{N}[C_{77}(X_1^N|Y_1^N) \log N - C_{77}(X_1^N) \log N] \quad (8)$$

Decide that $f_{ZMM1}(\mathbf{X}, \mathbf{Y}) = 1$ (i.e. Q and P are identical) if $d_{ZMM1}(Z_1^N||X_1^N) \leq \epsilon$. Otherwise, set $f_{ZMM1}(\mathbf{X}, \mathbf{Y}) = 0$ (i.e. decide that Q is different from P). The following Lemma states that the classifier f_{ZMM1} that is described above is asymptotically optimal over every finite class $\hat{M}_\ell \in M_\ell$ of processes.

Note that the ZMM measure that is used in [1] is slightly different, namely:

$$d_{ZMM}(X_1^N|Y_1^N) = \frac{1}{N}[C_{77}(X_1^N|Y_1^N) \log N - C_{78}(X_1^N) \log N]$$

Lemma 1 *Applying f_{ZMM1} to any finite class $\hat{M} \in M$ of processes yields,*

$$\limsup_{N \rightarrow \infty} \lambda(\hat{M}) = 0$$

Proof of Lemma 1: Lemma 1 above follows directly from [1] for the case where \hat{M} is restricted to be a finite class of finite- order Markov processes with positive transitions. However, here we deal with the more general case where \hat{M} may be any finite subset of the vanishing-memory collection M . This calls for a slight variations in the proofs that appear in [1].

Consider the vector $X_{-k_1}^{k_2}$ where k_1, k_2 are two arbitrary positive integers. Then, for any probability measure $P(\cdot) \in M_\ell$ we have, by definition (positive transition property and strong mixing),

$$P(X_{-k_1}^{k_2}) = P(X_{-k_1}^k, X_{k+1}^0, X_1^{k_2}) > \frac{1}{\beta} P(X_{-k_1}^{-k})(\delta)^k P(X_1^{k_2}) \geq \frac{1}{\beta} P(X_{k_1}^0) P(X_1^{k_2}) \delta^k$$

and,

$$\begin{aligned} P(X_{-k_1}^{k_2}) &\leq \beta P(X_{-k_1}^{-k})(1 - \delta)^k P(X_1^{k_2}) \\ &\leq \beta P(X_{k_1}^{-k}) P(X_1^{k_2}) (1 - \delta)^{2k} \frac{1}{\delta^k} \leq \beta P(X_{-k_1}^0) P(X_1^{k_2}) \frac{1}{\delta^k} \end{aligned}$$

Re-derive Eq.(23), Eq.(26) and Eq.(32) in [1] for the more general "strong mixing" model that

is adopted here (replacing ℓ in Eq.(32) by k and n by N) yields:

$$E\bar{\delta}(z^L) \leq [1 - \frac{1}{B}\bar{\delta}^k N^{-(1-\mu)}](\frac{N}{L} - 1) \quad (9)$$

Eq.(9) above replaces Eq.(23) in [1].

Also, the following equation replaces Eq.(26) in [1],

$$-\log P(\mathbf{z}) \geq (1 - \mu)(\bar{c} - 1) \log N - \bar{c}(k \log \frac{1}{\delta} + \log B) \quad (10)$$

. where \bar{c} is defined in [1]. In a similar way, Eq.(32) in [1] is replaced by,

$$-\log P(\mathbf{z}) \leq -\sum_{i=1}^{\hat{c}-1} \log P(z^{L_i}) + \hat{c}k[\log \frac{1}{\delta} + \frac{1}{k} \log B] \quad (11)$$

where \hat{c} is defined in [1]. Thus, Eq.(28) and Eq.(35) in [1] remain valid.

The proof then follows from steps that are similar to the steps that leads to part a) and part b) of Theorem 1 in [1], and by the fact that $\lim_{N \rightarrow \infty} [-\frac{1}{N} \log P(\mathbf{X}) - \frac{1}{N} C_{77}(\mathbf{X}) \log N] = 0$, almost surely.

After establishing the *asymptotic* optimality of f_{ZMM1} classifier we proceed to demonstrate it's *essential*-optimality, as defined above. Consider again the the class of processes \hat{M} that was used in the proof of Theorem 6 in [1] and in the proof of Theorem 1 above and let $N_0 \geq 2^{\ell(R+\epsilon)}$. Then,

Theorem 2 *For some small positive number ϵ and for a large enough ℓ*

$$\lambda(\hat{M}_\ell) \leq \max_{P \in \hat{M}_\ell} P_r[\mathbf{X}, \mathbf{Y} : \frac{1}{N_0} [C_{77}(\mathbf{X}|\mathbf{Y}) \log N_0 - C_{77}(\mathbf{X}) \log N_0] \leq \epsilon \text{ for some } Q \neq P \text{ or}$$

$$\frac{1}{N} [C_{77}(\mathbf{X}|\mathbf{Y}) \log N_0 - C_{77}(\mathbf{X}) \log N_0 > \epsilon \text{ and } Q = P] \leq O(\frac{1}{\ell \log \ell})$$

where $Q, P \in \hat{M}$.

Proof of Theorem 2:

Parse \mathbf{X} to generate a concatenation of ℓ -vectors (except, perhaps of the last vector in the generated concatenation), namely

$$\mathbf{X} = X_1^\ell, X_{\ell+1}^{2\ell}, \dots, X_{i\ell+1}^{(i+1)\ell}, \dots, X_{n\ell+1}^{N_0} \quad (12)$$

where n is an integer satisfying $N_0 - 1 < n\ell \leq N_0$. Define

$$\begin{aligned} \delta(X^\ell, \mathbf{Y}) &= 1 \text{ if } X^\ell = Y_{i-\ell+1}^i \text{ for some } i \in [0, N_0 - \ell + 1] \text{ and} \\ \delta(X^\ell, \mathbf{Y}) &= 0. \text{ otherwise.} \end{aligned} \quad (13)$$

Now, by Eq.(A12) in [2]

$$P_i(X_1^\ell | X_1^\ell \in A_j) \leq P_r(|W| \geq \frac{\beta_0}{2} \frac{1}{|A_j|}) \leq 2^{-\ell(R_0 - c(\beta_0, \delta))} \quad (14)$$

where $c(\beta_0, \delta, \ell)$ is defined in Eq. (A12b) in [2].

Thus, by the union bound, for any $X_\ell \in A_j; j \neq i$

$$P_r(\delta(X_1^\ell, \mathbf{Y}) = 1) \leq N_0 2^{-\ell(R_0 - c(\beta_0, \delta))} \quad (15)$$

Since $N_0 = 2^{\ell(R+\epsilon)}$, and setting $R_0 = R - \frac{\epsilon}{2}$ yields,

$$E \sum_{i=1}^{\frac{n}{\ell}} \bigcup_{X_1^\ell \in [0,1]^\ell} \delta(X_{i\ell+1}^{(i+1)\ell}, \mathbf{Y}) \leq \frac{n}{\ell} 2^{-\ell(c(\beta_0, \delta) - \frac{\epsilon}{2})} \quad (16)$$

where $E(\cdot)$ denotes expectation relative to $P_i(\cdot)$.

Also, since at least one LZ77 phrase must either begin or end in any ℓ -vector for which

$\delta(X_1^\ell, \mathbf{Y}) = 0$, and by the Markov inequality it follows that

$$P_r[C_{77}(\mathbf{X}|\mathbf{Y}) - 1, \text{ for some } Q \neq P; Q, P \in \hat{M}_\ell] \leq \frac{N_0}{\ell} [1 - 2^{-\frac{1}{2}\ell(c(\beta_0, \delta) - \frac{\epsilon}{2})}] \leq 2^{-\frac{1}{2}\ell(c(\beta_0, \delta) - \frac{\epsilon}{2})} \quad (17)$$

Next, $C_{77}(\mathbf{X})$ is evaluated for any $Q \in \hat{M}_\ell$. By construction of \hat{M}_ℓ above, the number of LZ77 phrases in each of the consecutive $\nu\ell + \nu'\ell$ -letters substrings in \mathbf{X} is no more than $O(\frac{\ell}{\log \ell})$, almost all of which appear in the first ℓ -vector that then repeats itself.

Hence,

$$\frac{1}{N_0}C_{77}(\mathbf{X}) \leq O\left(\frac{1}{\nu \log \ell}\right) \quad (18)$$

Therefore, by Eqs (17) and (18)

$$P_r[\mathbf{X}, \mathbf{Y} : \frac{\log N_0}{N_0}[C_{77}(\mathbf{X}|\mathbf{Y}) - C_{77}(\mathbf{X})] \leq \epsilon \text{ for some } Q \neq P; Q, P \in \hat{M}_\ell] \leq 2^{-\frac{1}{2}\ell(c(\beta_0, \delta)) - \frac{\epsilon}{2}} \quad (19)$$

for any $0 < \epsilon < 1$ and a large enough ℓ .

The last step of the proof of Theorem 2 demonstrates that the classification error vanishes also in the case where $Q = P$.

Let $\gamma_0 = 1 - \delta$, and let $\nu_0 > 1$ satisfy $\gamma_0 \nu_0 < 1$. Then, following the derivation of Eq.(67) in [2], for any $Z : P(Z) \geq \frac{1}{N_0 2^{\epsilon_0 \ell}}$

$$P_r(\delta(Z, \mathbf{Y}) = 0) \leq \beta(nu_0 \gamma_0)^\ell + 2^{-\epsilon_0 \ell} = 2^{-k\ell}; \text{ for some } k > 0 \quad (20)$$

Now, by construction and by Eq. (A10) in [2], each process in \hat{M}_ℓ consists of statistically independent "vectors" of length $\nu\ell + \nu'$ bits, where the probability of each such vector is lower-bounded by:

$$P(X_1^{\nu\ell + \nu'}) \geq 2^{-\ell R} \quad (21)$$

for large enough ℓ , with probability $1 - P_r(|W| \geq \frac{\beta_0}{2} \frac{1}{|A_j|}) \geq 1 - 2^{-\ell(R_0 + c(\beta_0, \delta))}$

Thus, by Eqs.(20) and(21), and since no more than one LZ77 phrases either starts or ends in any vector Z in \mathbf{X} for which $\delta(Z, \mathbf{Y}) = 0$ and since no $(\nu\ell + \nu')$ -vector contains more than $O(\frac{(\nu+1)\ell}{\log \nu\ell})$ LZ77 phrases, it follows that

$$E[C_{77}(\mathbf{X}|\mathbf{Y}); Q = P; P \in \hat{M}_\ell] \leq \frac{N_0}{\nu\ell} [1 + (2^{-\ell(R_0 + c(\beta_0, \delta))} + 2^{-k\ell}) O(\frac{\nu\ell}{\log \nu\ell})] \quad (22)$$

Eqs. (19) and (22) and the Markov inequality, and choosing large enough ν and ℓ lead to the completion of the proof of Theorem 2.

An Empirical Statistics Classifier (ESC) is not essentially optimal

Let $N = 2^{\ell(R-\epsilon)}$ and let the empirical measures $\hat{P}_{\mathbf{X}}(Z_1^n)$ and $\hat{Q}_{\mathbf{Y}}(Z_1^n)$ be based on the recurrence time of Z_1^n in \mathbf{X} and in \mathbf{Y} utilizing Kac lemma as in [2] where $\frac{1}{\delta^n} > P_{\mathbf{X}}(Z_1^n) > \delta^n$ and $\frac{1}{\delta^n} > Q_{\mathbf{Y}}(Z_1^n) > \delta^n$.

Then , by [2, Eq.(68)], there exists a small positive $\delta_0 \ll 1$ such that,

$$P_r[\log \hat{P}_{X_1^N}(Z_1^n) - \log P(Z_1^n) > n\epsilon_0 \text{ for some } Z_1^n \in \mathbf{A}^n] \leq (2\beta + 1)2^{n \log A} 2^{-n\epsilon_0} < 2^{-n\hat{c}(\delta_0, \epsilon_0)}$$

Thus, by the Markov inequality

$$P_r\left[\left|d(X_1^{N_0}, Y_1^{N_0}) - \frac{1}{N} \sum_{i=0}^{T-1} \log \frac{P(X_{i+1}^{(i+1)n})}{Q((X_{i+1}^{(i+1)n}))}\right| > \epsilon_0 + \frac{1}{\delta} 2^{-n \frac{\hat{c}(\delta_0, \epsilon_0)}{2}}\right] \leq 2^{-n \frac{\hat{c}(\delta_0, \epsilon_0)}{2}}$$

By Theorem 1 above, $\lambda(\hat{M}_\ell) \approx 1$

If the length of the test sequences is increased from $N = 2^{\ell(R-\epsilon)}$ to $N^* = 2^{\ell(R+\epsilon)}$, $n = \delta_0 \log N$ is only slightly increased, $n^* = \delta_0 \log N^*$ and therefore $n^* - n = \delta_0 2\epsilon \ell$

By the δ -positive transition property ,

$$\frac{1}{n^*} \left| \log \frac{P(X_1^{n^*})}{Q(X_1^{n^*})} - \log \frac{P(X_1^n)}{Q(X_1^n)} \right| \leq \frac{1}{n_0} \delta_0 2\epsilon \ell$$

Thus, no abrupt change in the value of $d(X_1^N, Y_1^N)$ if the length is increased from N to N^* .

Hence, $\lambda(\hat{M}_\ell) \approx 1$ even if the length of the sequences is increased to N_0 , for any Δ for large ℓ .

Also, it follows from the definition of the class M that

$$\left| \frac{1}{N} \sum_{i=0}^{T-1} \log \frac{P(X_{i+1}^{(i+1)n})}{Q((X_{i+1}^{(i+1)n}))} - \frac{1}{N} \log \frac{P(X_1^N)}{Q((X_1^N))} \right|$$

vanishes as N gets large and hence the ESC is asymptotically optimal.

However, as demonstrated above, it is NOT essentially optimal.

Section 2: A Variable length Fidelity Function F_{VL}

Let $L_{i,N,P}(X_i^N) = \max_{j=1}^{L_{max}} [j : P(X_1^j) \geq \frac{1}{N}]$ where $L_{max} = O(\log N)$.

Define:

$$F_{N,VL}(P, Q) = \frac{\log N}{E_P[L_{1,N,Q}(X_1^N)]} - \frac{\log N}{E_P[L_{1,N,P}(X_1^N)]} \quad (23)$$

Observe that due to the positive transition property of the class M , $L_{i,N,P}(X_i^{i+L_{max}}) \geq \frac{\log N}{\log \frac{1}{\delta}}$ and increases monotonically with N .

Thus, there exists some large enough $L_{min}(N_0)$ such that for $N \gg N_0$, each $L_{1,N,P}(X_1^N)$ -vector consists of a large number of $L_{min}(N_0)$ -vectors with a guard-space of k_0 letters in between any such two consecutive vectors, that, by the vanishing memory property, are approximately independent from each other. Thus, with high probability, each $L_{1,N,P}(X_1^N)$ vector consists of about the same composition of $L_{min}(N_0)$ vectors.

It then follows by the central limit theorem that almost surely, relative to the P measure,

$$\lim_{N_0 \rightarrow \infty} \lim_{N \rightarrow \infty} \left[\frac{\log N}{L_{1,N,Q}(X_1^N)} - \frac{\log N}{E_P[L_{1,N,Q}(X_1^N)]} \right] = 0.$$

Similarly, almost surely, relative to the P measure,

$$\lim_{N_0 \rightarrow \infty} \lim_{N' \rightarrow \infty} \left[\frac{\log N'}{L_{1,N',Q}(X_1^{N'})} - \frac{\log N'}{E_P[L_{1,N',Q}(X_1^{N'})]} \right] = 0.$$

Setting N' so as to make $E_P[L_{1,N',P}] = E_P[L_{1,N,Q}]$ leads to the conclusion that

$F_{N,VL}(P, Q)$ tends to $D(P||Q)$ almost surely and hence $F_{N,VL}(P, Q) \in \mathbf{F}$ as required.

An optimal universal F_{VL} -classifier

Consider the class \hat{M} which was used in the proof of Theorem 1 above.

It follows by Lemma A1 in [2] that just like $D_N(Q||P)$, $F_{N,VL}(Q, P)$ can be made arbitrarily large by selecting δ to be small enough. It then follows that Theorem 1 holds for the variable length fidelity measure as well, and therefore if $N \leq N_0^{-\epsilon\ell}$, the probability of classification error must be

close to one for any universal classifier.

A universal classification algorithm that yields a negligible classification error for any Q, P pair for which $F_{N,VL}(Q.P) \geq \Delta > 0$ if $N \geq N_0 2^{\epsilon \ell}$ is now introduced. Similar to the ZMM in [1], it is based on cross-parsing..

Let

$$L_{i,N}(\mathbf{X}) = \max_{j=1}^{L_{max}} [j : X_{i+1}^{i+j} = X_t^{t+j} \text{ for some } 1 \leq t \leq \frac{N}{2} + 1]$$

Let M be a positive integer satisfying $M = N^{1-\frac{\epsilon}{2}\ell}$ where ϵ is an arbitrary small positive number.

$$L_{i,N,M}(\mathbf{X}|\mathbf{Y}) = \max_{j=1}^{L_{max}} [j : X_{i+1}^{i+j} = Y_t^{t+j} \text{ for some } \frac{kN}{M} + 1 \leq t \leq \frac{(k+1)N}{M} - j \text{ and every } 1 \leq k \leq \frac{M}{N} - 1]$$

and let

$$\tilde{L}_{N,Q,M}(\mathbf{X}|\mathbf{Y}) = \frac{1}{N - L_{max}} \sum_{i=1}^{N-L_{max}} L_{i,N,M}(\mathbf{X}|\mathbf{Y}) \quad (24)$$

and,

$$\tilde{L}_{N,P}(\mathbf{X}) = \frac{1}{\frac{N}{2} - L_{max}} \sum_{i=1}^{\frac{N}{2} - L_{max}} L_{i,N}(\mathbf{X}) \quad (25)$$

Set $f_c(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ (i.e. $Q \neq P$) if:

$$\frac{\log N}{\tilde{L}_{N,Q,M}(\mathbf{X}|\mathbf{Y})} - \frac{\log N}{\tilde{L}_{N,P}(\mathbf{X})} \geq \Delta + \epsilon$$

and $f_c(\mathbf{X}, \mathbf{Y}) = \mathbf{1}$ (i.e. $Q = P$) if:

$$\frac{\log N}{\tilde{L}_{N,Q,M}(\mathbf{X}|\mathbf{Y})} - \frac{\log N}{\tilde{L}_{N,P}(\mathbf{X})} \leq \epsilon \text{ for a preset small positive number } \epsilon \ll \Delta.$$

Lemma 2 For any arbitrary small positive δ , there exists an ℓ_0 such that for any $\ell > \ell_0$

$$\sup_{Q \in M} P_r[Q : |L_{i,N}(\mathbf{X}|\mathbf{Y}) - \mathbf{L}_{i,N,Q}(\mathbf{X}_1^N)| \geq \delta] \leq \mathbf{L}_{\max} \mathbf{A}_{\max}^L 2^{-\mathbf{Mc}(\ell, \mathbf{k}_0, \delta)\ell} \leq \delta$$

for $N = N_0 2^{\epsilon \ell}$ and $M = N_0 2^{-\frac{\epsilon}{2}\ell}$

and therefore,

$$\sup_{Q \in M} P_r[Q : |\tilde{L}_{Q,N,M}(\mathbf{X}|\mathbf{Y}) - \frac{1}{N - L_{max}} \sum_{i=1}^{N-L_{max}} L_{i,N,Q}(X_1^N)| \geq \delta] \leq \delta$$

Also,

$$|\tilde{L}_{P,N}(\mathbf{X}) - \frac{1}{\frac{N}{Q} - L_{max}} \sum_{i=1}^{\frac{N}{2} - L_{max}} L_{i,N,P}(X_1^N)| \geq \delta] \leq \delta$$

The proof follows directly from Kac's Lemma and the properties of the class M (see Eq (68a) in [2].

Lemma 3 *Let $N = N_0 2^{\epsilon \ell}$. Then,*

1)

$$P_r[|\tilde{L}_{N,Q,M}(\mathbf{X}|\mathbf{Y}) - E_P L_{1,N,Q}(\mathbf{X}|\mathbf{Y})| \geq \epsilon] \leq 2^{-c(k_0, \beta) 2^{\epsilon \ell}}$$

2)

$$P_r[|\tilde{L}_{N,P}(\mathbf{X}) - E_P L_{1,N,P}(\mathbf{X})| \geq \epsilon] \leq 2^{-c(k_0, \beta) N_0^\epsilon}$$

for some $c(k_0, \beta) > 0$ where $\beta < 2^{\frac{\epsilon}{2}}$.

Proof of Lemma 3: Parse \mathbf{X} into consecutive substrings of $n_0 + k_0 + L_{max}$ letters each, where $n_0 = K(k_0 + L_{max}; K \gg 1)$. Observe that by the vanishing memory property, the successive blocks of n_0 letters are "almost" independent since they are separated by a guard space of $k_0 + L_{max}$ letters and are governed, up to a factor of β^K , by a K -th product memoryless probability measure of n_0 -vectors.

Also observe that

$$\tilde{L}_{N,Q}(\mathbf{X}|\mathbf{Y}) = \tilde{L}_{N,Q}(X_1^{n_0}|\mathbf{Y}) + \tilde{L}_{N,Q}(X_{n_0+1}^{n_0+k_0}|\mathbf{Y})$$

and,

$$\tilde{L}_{N,P}(\mathbf{X}) = \tilde{L}_{N,P}(X_1^{n_0}) + \tilde{L}_{N,P}(X_{n_0+1}^{n_0+k_0}).$$

But,

$$0 \leq \tilde{L}_{N.Q}(X_{n_0+1}^{n_0+k_0} | \mathbf{Y}) \leq L_{max} k_0 \text{ and,}$$

$$0 \leq \tilde{L}_{N.P}(X_{n_0+1}^{n_0+k_0}) \leq L_{max} k_0$$

Setting $n_0 = \frac{2L_{max}k_0}{\epsilon}$, $N_0 = n_0 + k_0 + L_{max}$, and $\beta < 2^{\frac{\epsilon}{2}}$ leads to Lemma 1 by applying the Chernoff bound for sums of i.i.d bounded random variables. This leads to Lemma 3 and therefore, to the F_{VL} -optimality of the proposed algorithm, which has a computational complexity that is proportional to $N \log N$.

In conclusion, it should be pointed out that by slightly modifying the ZMM algorithm in section 1 above by replacing $C_{77}(X_1^N)$ with $2C_{77}(X_1^{\frac{N}{2}} | X_{\frac{N}{2}+1}^N)$, and by changing the cross-parsing procedure that led to $C_{77}(X_1^N | Y_1^N)$, where each phrase now is the longest incoming string of the yet unparsed letters that appears in *all* of the $\frac{N}{M}$ M sub-blocks in \mathbf{Y} , where $M = N_0 2^{\frac{\epsilon}{2}\ell}$, one gets a universal classification algorithm, which by the same arguments that led to Lemma 2 and Lemma 3 above can be shown to be F_{VL} -optimal as well (at least for $\Delta > 4$). However, following [4] where $\frac{\log N}{\tilde{L}_{N.P}(\mathbf{X})}$ was demonstrated to be an efficient entropy estimator, which was shown to converge to the entropy faster than one based on LZ77, it appears that since the algorithm above utilizes $O(N)$ data points rather than $O(\frac{N}{\log N})$ in the modified ZMM algorithm case, the latter may yield a classification error probability that converges to zero at a slower pace.

Acknowledgement

Helpful comments by Neri Merhav are acknowledged with thanks.

References

- [1] J.Ziv, N.Merhav “A measure of relative entropy between individual sequences with application to universal classification” *IEEE Trans. Inf. Theory*, vol. IT-39, no. 4, pp. 1270–1279, July 1993.

- [2] A.D. Wyner, J. Ziv, "Classification with finite memory", *IEEE Trans. Inf. Theory*, vol. IT-42, no. 2, pp. 337-347, March 1996.
- [3] J. Ziv, "Classification with finite memory revisited" *IEEE Trans. Inf. Theory*, Vol. IT-53, Issue 12, Page(s):4413 - 4421, Dec. 2007.
- [4] J. Ziv, A. Lempel "A Universal Algorithm for Sequential Data Compression" *IEEE Trans. Inf. Theory*, vol. IT-23, no. 3, pp. 337-343 May 1977
- [5]] Kontoyiannis, I. , P.H. Algoet, Yu. M. Suhov and A.J. Wyner. Non-parametric entropy estimation for stationary processes and random fields, with applications to English text .*IEEE Transactions Information Theory*. Vol. IT-44, pp. 1319 - 1327, May, 1998.